# Offensive language without offensive words (OLWOW)

Klenner, Manfred

# Offensive Language without Offensive Words (OLWOW)

**Manfred Klenner**
Institute of Computational Linguistics
University of Zurich
Switzerland
klenner@cl.uzh.ch

## Abstract

In our contribution, we have applied stance analysis in order to identify offensive discourse. This gives us access to the pros and cons of the writer of some tweets and reveals his/her role framing of the discourse referents. We also semi-automatically augmented our polarity lexicon with a new type of polarity labels, namely P for profanity. Starting from seed words, we derived new entries on the basis of word embeddings. Our approach also focuses on offensive language without offensive words (OLWOW) and discusses strategies to cope with it.

## 1 Introduction

The GermEval Task 2018 deals with offensive language. The training material are about 5,000 German tweets classified (task I) as offensive (label OFFENSE) compared to not offensive (label OTHER). Task II further specifies offensive language as profanity, abuse or insult. According to the annotation guidelines, profanity indicates the use of indecent, nasty or vulgar vocabulary, while insult and abuse moreover are given, if such words are used to characterise the attributes of a person (INSULT) or to assign a negatively connotated social class to a person (ABUSE). See the following examples insult (ex. 1), abuse (ex. 2) and profanity (ex .3).

**ex. 1.** Merkel ist die grösste Versagerin der Weltgeschichte !!! (Merkel is the biggest loser in world history)

**ex. 2.** Clinton - Der Antichrist (Clinton -the antichrist)

**ex. 3.** Ist zum kotzen (it sucks)

After a couple of attempts to predict the annotations of the gold standard, the author of this paper is convinced that this annotation task was not trivial. I still believe that the annotations of a couple of sentences are debatable.

About one third of the data is classified as offensive language, where abuse is the majority class (about 1,000 tweets), followed by insult (600 tweets) and complemented by a small profanity sample (70 examples). The majority baseline for task I - assigning OTHER - yields an accuracy of 66.3%.

A glimpse at the data reveals that - as expected - the vocabulary being used is the central indicator of offensive language. This seems to prompt for a lexicon-based solution, although the resulting task then is to deal with unknown words. Especially compounds are a very flexible means to create new words in German. But the number of vulgar words is large, anyway, so a mechanism to induce such words is needed. Word embeddings might help. Thus deep learning comes into play. However, we were not so much interested in the best performing black box, but wanted to find out whether our stance analysis system based on a purely symbolic computation could be of any use.

## 2 Resources

The organisers provided a couple of resources, among others German word embeddings, but also lexicons with e.g. German swearwords. We only integrated one resource, the swearword lexicon. We did it semi-automatically. First, we determined the frequency of each word in a corpus of Facebook posts from a German right-wing party. Then we had a look at the most frequent words and kept 300 of them. We added these words to our polarity lexicon for German, comprising 6,800 nouns and adjectives classified as positive or negative in one of three dimension, namely, the dimension of emotion, moral or appreciation (following the distinction of the appraisal theory, cf. (Martin and White., 2005)). We also have specified a verb lex-

icon comprising 1.100 verbs, where a verb might have various frames indicating the syntactic frame of the verb and whether the verb has a polar effect on its arguments (positive or negative). For example, the verb *anpöbeln* (to accost sb, to molest sb) casts a negative effect on its agent role (which is the source) and on its patient role (the target). Also a negative relation (con) between source and target is assigned (given that the verb is being affirmatively used). This forms the basis of our system for stance analysis. We also assigned verb specific polar roles to source and target. For instance, the patient of the target role of *verleumden* (to slander, slur, vilify) is said to be a victim while the source takes the role of a villain. We call the assignment of polar roles to discourse referents *role framing*, since it conceptualizes a referent in a specific way. It represents the writer perspective. It indirectly indicates the writer's stance towards the referents: he/she is against the villain but in favour of the victims.

Although we are dealing with tweets, we applied an ordinary dependency parser (Sennrich et al., 2013). We just stripped hash tags, emoticons and other social media language noise.

## 3 Qualitative Analysis

Although it is rather evident that - for a good performance - a subsymbolic approach would be well suited (either character level n-grams or deep learning), we pursued another approach. Our goal was to find out, whether our system for stance analysis could help to understand the problem and help to solve the task. The idea was to first identify the proponents and opponents of the writer of the (offensive) tweets and then to look for polar relations where e.g. a proponent of the writer received a negative effect, or the opponent of the writer received a positive effect. We thought that such constellations might bear conflict potential which - in the best case - would be the yeast of offensive language usage. Very soon we realised that we still had to deal with vocabulary gaps, since most of the time offensive language is based on the usage of offensive words. Actually, our hope was that we were able to identify exactly those cases of (implicit) offensive language that are not indicated by offensive words. We give a couple of examples (cf. examples ex.4 to ex.6).

**ex. 4.** Das deutsche Volk wird unaufhörlich belogen! (The German people are constantly being lied to!)

**ex. 5.** Merkel muss weg. (Merkel has to go.)

**ex. 6.** Sie warnen vor Nazis und führen deren Methoden der Bücherverbrennung und Meinungsunterdrückung ein. (They warn against Nazis and introduce their methods of burning books and suppressing opinion.)

Example ex.4: our system derives that *Volk* is a victim (after passive normalization), since the target of *belügen* (lie to) in an affirmatively used sentence is a victim (the source is a villain, but no source is given here). Example ex.5: a negative effect applies to *Merkel* stemming from *wegmüssen*. We are not able to deal with example ex.6 at the moment. Although a *con* relation from the source (they) to the target (Nazis) is derived, and although we were able to deduce a positive effect on *they*[1] the implicit contrast with the second conjunct (following the "and") is beyond the current capabilities of our system.

These sentences contain no offensive words, but are annotated as offensive language. How to deal with these sophisticated examples?

## 4 Model Based on Lexicon

We trained a word2vec model on the basis of three Swiss newspapers (NZZ, Tagesanzeiger, Blick). In order to find new examples of offensive words, we manually specified a seed lexicon comprising 20 words. On the basis of the gensim module, we then generated for each seed word the 25 closed neighbors and manually removed false positives. After three rounds, we ended up with 275 entries.

We randomly extracted 500 tweets from the training set as a preliminary test set and carried out several experiments with the full polarity lexicon and subversions of it. This revealed that the precision was ok, but recall was a bit low. Next we calculated the correlation between words of the training set and the offensive class. This gave better results. The precision of OFFENSE was 61.41%, recall was 69.32%, f1 was 65.12% and accuracy was 75.80 %. We took this as our starting point. We now turn to a more detailed description of our approach.

Rather quickly it became clear that some words are very good indicators of offensive language. For instance, the word *Scheiss* (shit) perfectly indicates the class OFFENSE. We thus decided to simply predict the class of a tweet on the basis of these words.

[1] A negative effect on a negative target gives a positive effect on the source of such a situation.

We estimated the probability of an offensive class given a word W with the following approximation:

$$P(OFFENSE|W) \approx \frac{\#(W, OFFENSE)}{\#W}$$

This is: the number of times OFFENSE is the label of a tweet that includes word W divided by the number of times word W occurs in the training corpus. We kept those words that have a probability above 0.75 and of a frequency in the corpus above a THRESHOLD which is 2 for words not in the polarity lexicon and 0 for words from the polarity lexicon. We call this filter the *word indicator filter*(it comprises 508 words) and used it as a classifier in the following way.:

$$P(OFFENSE|TWEET) = 1 \; if$$
$$\exists\, W \in TWEET : P(OFFENSE|W) > 0.75$$
$$\wedge\; freq(W, CORPUS) > THRESHOLD$$

If a tweet contains one word of the filter it is classified as OFFENSE. There are other filters: verb related filters (see next section) and an exclamation mark filter. Those tweets that pass all filters are classified as OTHER.

There are a couple of possible correlations one could take into consideration and a machine learning tool could do this much better than a manual engineer. However, since we were not so much interested in exploiting indicators that are language independent (like the number of hash tags being used, capital letter usage etc.), but rather in the language specific means, we have not undertaken a detailed analysis on that level. The only exception are exclamation marks. If a tweet contains more than two successive exclamation marks, it is classified as offensive. This is the *exclamation mark filter*. Let us now turn to our stance-based filters.

## 5 Model Based on Stance Detection

Our stance analysis is verb-based (Klenner et al., 2016). It only triggers if a model verb with the right syntactic frame (and sometimes further lexical restrictions) is present. Then, dependent on the verb and its affirmative status (negated or not), role framing, i.e. the assignment of polar roles occurs and a polar relation (pro or con) is established from the source towards the target. The main polar roles are *victim, villain, benefactor, beneficiary, pos_actor, neg_actor, neg_affected, pos_affected*. They are associated with the *source* and *target* (cf. (Wiegand et al., 2016)) of a verb. The source marks the semantic roles of the initiator of the positive or negative relation that a verb expresses towards the target. For instance the verb *to cheat*: the direct object (patient) is the target as well as a victim and the source is the (logical) subject (agent) and it is modelled as a villain (since *to cheat* is morally negative). Our stance model claims that role framing, the assignment of polar roles, reveals the writer perspective, since if the writer conceptualises someone as e.g. a villain, he/she is against this referent. Finding the targets of the stance of the author, thus, boils down to analyse his/her role framing. If the proponents and opponents of the writer are known, we can start to infer additional proponents and opponents of his/her. For instance, if someone is in favour of a proponent of the writer, then this person becomes a candidate proponent of the writer. So if the EU is a proponent of mine and you praises the EU, you might be a proponent of mine. We do not need the full-fledged capabilities of our stance system. We wanted to explore the idea that we were able to identify offensive language, namely the cases where no offensive vocabulary is present.

But the first question was: is our approach comprising 1,100 verbs and about 1,700 different frames plagued by sparseness? In 827 of the 3532 sentences from the test set it triggered. This is 23.4 % of all sentences (for the training set it is 25.38%). This is not too sparse. This gave us 818 polar roles and 176 pro (73) and con (103) relations, altogether 994 assignments. The first step in stance analysis is to find the targets of the writer: who is he/she against or in favour of? We just took those referents conceptualized as villains and neg_actors: $\lambda x : villain(x) \vee neg\_actor(x)$. The result comprises *SPD (a political party), Mob (mob), Salafisten (Salafists), Einwanderer (immigrants), Lügenpresse (lie press), Merkel (German chancelor), Allah (Allah)*. Obviously, the (some) writers are against these referents. And who are the victims? We get (among others): Volk (people), Jude (jew), Planet (planet), Polizist (cop), Deutschland (Germany), Sicherheit (safety), Kind (child), Frauen (women).

Are there correlations we could exploit: e.g. between role framing and the class OFFENSE? We run quite a number of tests. E.g. we determined the probability $P(OFFENSE|villain) = 0.66$, but the are only 35 cases. Other examples are: $P(OFFENSE|neg\_actor) = 0.51$, $P(OFFENSE|victim) = 0.58$, $P(OFFENSE|pos\_actor) = 0.29$. That is,

pos_actor indicates OTHER with a probability of 71%. When it comes to pro and con relations, we got $P(OTHER|pro) = 0.73$ and $P(OTHER|con) = 0.60$. As we can see, a correlation between polar facts and binary classes (task 1) is given, but is not very striking. We use it as filters in our pipeline architecture.

The strongest filter is the word indicator filter. It is applied first. Tweets that do not pass it, are classified as OFFENSE. The rest runs through the filters: pro, pos_actor, villain and victim. Those who pass all filters are classified as OTHER. For our 500 sample test set derived from the training set, this gave us 61.41% precision and 69.32% recall.

## 6 Offensive Language without Offensive Words

In the training set there are a couple of examples of offensive language without offensive words (OLWOW). We created filters to identify such tweets. If a tweet triggers stance analysis and if a negative polar fact is derived, but none of its words are in our polarity lexicon, then this tweet is a candidate for an OLWOW. If, additionally, a negative polar fact hits an opponent of the writer, it is a candidate of OFFENSE. Here are three examples.

**ex. 7.** Es gibt bei uns keine Pressefreiheit mehr. (There is no longer a free press.)

**ex. 8.** Mal schauen wieviel Frauen dieses Jahr von illegalen Einwanderern vergewaltigt oder belästigt werden. (Let us see how many women get raped or harrased by illegal immigrants this year.)

**ex. 9.** Hier wird Vergewaltigung legalisiert! (Here, rape gets legalized!)

Example 7 and example 8 are annotated as ABUSE, while example 9 is a negative one, since it is annotated with OTHER. Our system is not able to deal with example 7 but correctly identifies example 8: women is classified as victim, immigrants as villain. Since immigrants are an element of the opponents and, in this sentence, are conceptualized as a villain (which is a negative effect), we are entitled to conclude that this tweet is offensive - although neither rape nor harass are offensive words. They denote aggressive events.

The concept of an OLWOW is demanding. According to the gold standard and our filters, 175 tweets are OLWOW tweets. However, if we require that the polar effect hits an opponent (our criteria for offensiveness), this is reduced to 9 cases. There are various reasons for the resulting sparseness: sometimes the parser has introduced wrong sentence boundaries, sometimes a pronoun occupies the polar role and we do not do coreference resolution, sometimes the cause for offensiveness is distributed over more than one sentence, etc. An example of a distributed representation is:

**ex. 10.** Wir haben Jerusalem vom Islam befreit und das heutige Banken System erfunden. Wer oder was sollte uns aufhalten. Merkel oder Maas etwa. Lachhaft. (We liberated Jerusalem from Islam and invented today's banking system. Who or what should stop us. Merkel or Maas? Ridiculous.)

As we can see, no offensive words are used and the abusive argumentation is distributed among 4 pieces. OLWOW annotations are also debatable since sometimes it is unclear whether we are talking about offensive language or just the freedom of speech. For instance example 7: is this not just an ordinary opinion?

We believe that OLWOW is an interesting and demanding research topic. Although we have explicated some conditions and discussed some ideas how to operationalize OLWOW detection, we could not make it fruitful for the task at hand because of sparseness.

## 7 Filter-based Model: GermEval Runs

We submitted three runs in the coarse-grained task setting.

We have filters that classify tweets as OFFENSE (word indicator, exclamation mark, neg_actor, villain, victim) and filters that classify tweets as OTHER (pro, pos_affected).

Run 1 (cluzh_coarse1.txt') includes the filters (in that sequence): pro, pos_affected, pos_actor, word indicator, exclamation mark. Run 2 (cluzh_coarse2.txt') includes the filters (in that sequence): word indicator, exlamation mark, neg_actor, villain and victim. Run 3 (cluzh_coarse1.txt') only includes the word indicator filter.

Tweets that pass all filters are classified as OTHER. We did not use the filters con, neg_affected, benefactor, beneficiary. Also the filters from the last section were not part of any submission because of the sparseness problem.

## 8 Conclusion

We presented a plain vocabulary-based approach to the detection of offensive language. We realised a

cascade of filters including verb-based ones coming from stance analysis. We also focussed on a particular interesting research topic that we named OLWOW, offensive language without offensive words (known as implicit offensive language). We discussed ideas how to cope with it, pointed out problems with the annotation process of OLWOW and presented of a couple of examples our stance analysis system is able to cope with. We could, however, not exploit this notion for our shared task runs due to the sparseness of trigger conditions. We have, however, gained some insights that we will explore in our future work.

**Acknowledgments**

I would like to thank Michi Amsler for interesting discussions, useful word embeddings and a list of nice swearwords.

## References

Manfred Klenner, Don Tuggener and Simon Clematide (2016). *Stance Detection in Facebook Posts of a German Right-wing Party.* In:LSDSem 2017/LSDSem Linking Models of Lexical, Sentential and Discourse-level Semantics, Valencia, 2017

J. R. Martin and P. R. R. White (2005). *Appraisal in English.* Palgrave, London, 2005

Rico Sennrich, Martin Volk and Gerold Schneider (2013). *Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis.* In: Proceedings of the International Conference Recent Advances in Natural Language Processing Hissar, Bulgaria, 2013

Michael Wiegand and Josef Ruppenhofer (2015). *Opinion Holder and Target Extraction based on the Induction of Verbal Categories.* Proceedings of the 19th Conference on Computational Natural Language Learning (CONLL) , Beijing, China, July 30-31, 2015